MATH1324 Assignment 3

Udeshika Dissanayake (S3400652)

May 27, 2018

Problem Statement

The objective of this study is to determine whether a human body circumference measurement could be used as a general indicator for human body fat percentage. A such body circumference measurement could then be used to predict the body fat percentage by establishing a simple linear formula. The study will further assess how well this linear formula perform to estimate the body fat percentage by comparing the predicted values against the real values. All the statistical computations have been performed in 'R Studio' package in this study.

A data-set of 252 people (160 male and 92 female) with their body fat percentages (Brozek method) and ten different body circumference measurements have been used in this study. The Source for the data-set: Roger W. Johnson. March 1996. Fitting Percentage of Body Fat to Simple Body Measurements. Journal of Statistics Education, Volume 4, Number 1.

Data Preparation

The data-set has been loaded in to R Studio and the outlier has been removed following the ranges, lower outliers < Q1-1.5 x IQR and
upper outliers < Q3+1.5 x IQR. The R packages "mosaic", "r markdown","kableExtra", "car" and "ggplot2" have been installed. The frequency distribution after the outlier removal is,

```
Sex
  male female  Total
   152     83    235
```

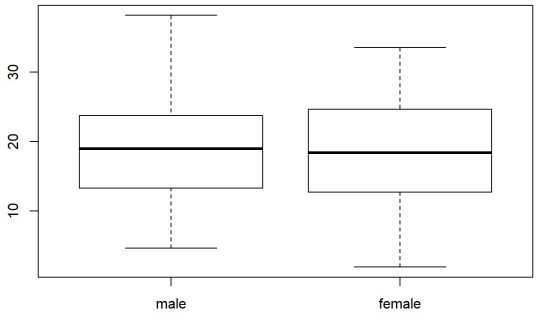
Part 1: Testing whether the mean body fat percentage for male and female are the same

Below code calculates the favorite descriptive statistics (i.e. mean, median, standard deviation, first and third quartile, minimum & maximum values, and number) of the body fat percentage for both male and female separately.

```
favstats(~BFP_Brozek|Sex,data=body)#Summary Statistics of body fat percentage by se
x
```

```
     Sex min    Q1 median     Q3  max     mean       sd   n missing
1   male 4.6 13.35   19.0 23.725 38.2 18.48158 7.167101 152       0
2 female 1.9 12.70   18.4 24.650 33.6 18.72892 7.415533  83       0
```

The difference between the mean values of male and female body fat percentages is, 18.482 - 18.729 = -0.247. The mean of the female body fat percentage is slightly higher than that of the male. It is worth performing two-sample t-test to determine whether this difference is statistically significant to consider.



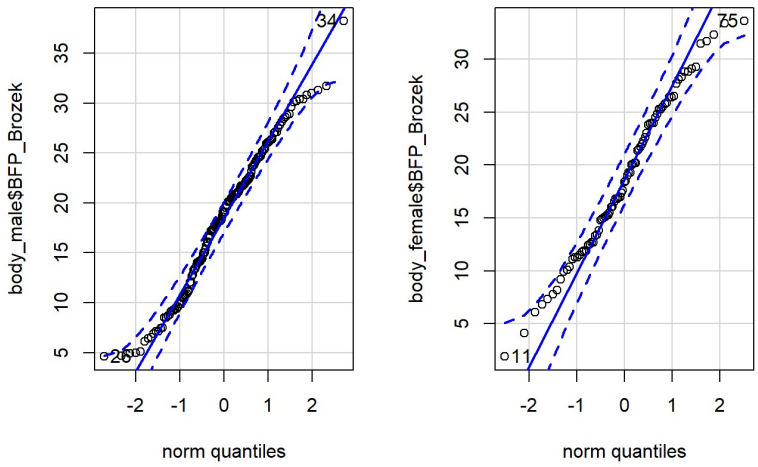
Two-sample T-test

The two-sample t-test is a powerful method to compare the means of two data populations if both follow the normal probability distribution. Further, the data points should be continuous, have equal variance, and independent of each other to qualify for this test.

Testing the assumption for Two-sample T-test

It is obvious that the male and female measurement are inherently independent.

By visually checking data placement in the Q-Q plot, the normality of the populations can be validated. Despite the data set is fairly large (> 30 point), Q-Q plot has been drawn for male and female body fat percentages as below:



As can be seen, the data points fall within the dashed lines that correspond to 95% CI for the normal quantiles. This proves that the two populations are normally distributed, hence qualify for two-sample t-test.

By using the p-value from the Levene's test, the variances of the two population could be tested for equality.

```
leveneTest(BFP_Brozek~Sex,data=body)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   1  0.3674  0.545
##       233
```

As can be seen above, the p-value for the Levene's test of equal variance for body fat percentage between males and females is p = 0.545. Since, p > 0.05, it could be assumed the variances are equal in above groups.

Hypothesis Tested

For comparing two means of the population groups, the basic null hypothesis is that the means are equal, whereas the alternative hypothesis is that the means are not equal,

$$H_0 : \mu_1 - \mu_2 = 0$$

$$H_A : \mu_1 - \mu_2 \neq 0$$

where μ_1 and μ_2 refer to the population means of male (group 1) and female (group 2) respectively.

Two-sample t-test - Assuming Equal Variance and a two-sided hypothesis test

Below R code performs the two-sample t-test for male and female body fat percentages assuming the variances are equal and two-sided.

```
t.test(BFP_Brozek~Sex,data=body, var.equal=TRUE, alternative="two.sided")
```

```

    Two Sample t-test

data:  BFP_Brozek by Sex
t = -0.24977, df = 233, p-value = 0.803
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.198306  1.703632
sample estimates:
  mean in group male mean in group female
            18.48158             18.72892
```

Following the p-value method, p = 0.803 > α = 0.05, the null hypothesis H_0 is fail to be rejected. This means there was not a statistically significant difference between the means.

Also, it is observed that H_0 = 0 is getting captured by the 95% confidence interval, 95% CI[-2.198 1.704]. Once again this suggests that there was no statistically significant difference between the means. In other words it can be statistically concluded that the means of body fat percentage for male and female are the same.

Part 2: Estimating 99% confidence interval for the mean body fat percentage in the popolation

The degree of uncertainty associated with fairly limited sample representing a large population can be assessed and quantified through the concept of confidence interval. Using the current sample data, it derives an interval, in to which the population parameters will keep on falling with percentage confidence if the random sampling procedure is repeatedly performed.

It is statistically conclusive from Part 1 study, that the male and female do have quite a similar distribution of body fat percentage, hence both male and female body fat percentage data samples could assumed to be same and could be treated equally for future assessments.

Summary Statistics of body fat percentage for the Data set

The favorite descriptive statistics are as follow, considering both male and female together:

```
favstats(~BFP_Brozek,data=body) #Summary Statistics of body fat percentage for the Sample
```

```
 min Q1 median Q3  max     mean       sd   n missing
 1.9 13     19 24 38.2 18.56894 7.240952 235       0
```

Assumptions for CI Assessments

In general, in order to perform CI assessment, the two population should have the same variance (i.e. homogeniarity of variance), populations should be normally distributed, and they should be independent. Each of these has been verified under the Part 1 study for the male and female. This means, if both male and female data points are combined, it should also be normally distributed, hence could safely perform the CI formula.

Further, since the data-set is fairly large (n > 30), it is well know that the CI formula works quite well even without normality assumption or normality checking.

Calculating 99% Confidence Interval

Below R code calculates the 99% CI for the subjective data-set:

```
confint(t.test(~BFP_Brozek,data=body,conf.level=0.99))
```

```
  mean of x    lower    upper level
1  18.56894 17.34225 19.79562  0.99
```

It is 99% confident that mean body fat percentage of random sampled of human would fall within the range of [17.342 19.796].

Part 3: Testing whether the average body fat percentage is less than 12.5

The one-sample t-test can be used to test whether there is enough evidence taken from a sample mean to suggest that the population mean is different to a assumed value under null hypothesis. The general assumptions used in for this method are data are normally distributed and the population standard deviation is unknown. These assumptions are reasonably proved in previous sections for body fat percentage data.

Hypothesis Tested

The null and alternative hypothesis are mentioned below for this study:

$$H_0 : \mu \geq 12.5$$

$$H_A : \mu < 12.5$$

The one-sample t-test (lower-tailed hypothesis test)

Below R code calculates the parameters for one-sample t-test analysis,

```
round(favstats(~BFP_Brozek,data=body),2)
```

```
 min Q1 median Q3  max  mean   sd   n missing
 1.9 13     19 24 38.2 18.57 7.24 235       0
```

```
t.test(~BFP_Brozek,data=body, mu=12.5, alternative="less")
```

```

    One Sample t-test

data:  BFP_Brozek
t = 12.848, df = 234, p-value = 1
alternative hypothesis: true mean is less than 12.5
95 percent confidence interval:
     -Inf 19.34897
sample estimates:
mean of x
 18.56894
```

As can be seen the 95% CI of the mean body fat percentage is [-Inf 19.349]. This means the mean value in the null hypothesis ($H_0 : \mu \geq 12.5$) is getting captured (at least partially) by CI. This concludes that the results of the Hypothesis test are not statistically significant, hence fail to reject the null hypothesis.

Also, using the p-value, since p = 1 > α , it is conclusive that null hypothesis is failed to be rejected.

This means the researches belief of average body fat percentage is less than 12.5 can not be supported with statistical significance.

Part 4: Finding the single best predictor of body fat percentage using the body circumference data

The simple linear regression model is used in this section to determine the correlation between a predictor variable (X - any body circumference measurement) and the dependent variable (Y - body fat percentage).

$$Y = \alpha + \beta X + \epsilon,$$

where α is the intercept, β is the slope, and ϵ is the random error with zero mean. The real data points are tested for best fit for linear regression model. The linear regression model needs below assumptions on the data-set to be true in order to maintain the integrity and validity of the model,

- Independence
- Linearity
- Normality of residuals
- Homoscedasticity

Each of these assumptions will be tested and validated for the best fit model in a later section in this report.

Hypothesis Tested

Below are the null and alternative hypothesis for the linear regression model, respectively:

H_0 : The data do not fit the linear regression model
H_A : The data fit the linear regression model

Linear Regression is fitted using lm() function

One body circumference measurement is randomly picked (Abdomen) and using below R code, the linear regression model curve is derived through estimating respective α and β values.

```
bfpAbdomenmodel <- lm(BFP_Brozek~Abdomen,data = body)#Linear Regression is fitted u
sing lm() function
msummary(bfpAbdomenmodel)
```

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -38.93848    2.82788  -13.77   <2e-16 ***
Abdomen       0.62641    0.03065   20.44   <2e-16 ***

Residual standard error: 4.342 on 233 degrees of freedom
Multiple R-squared:  0.6419,    Adjusted R-squared:  0.6404
F-statistic: 417.7 on 1 and 233 DF,  p-value: < 2.2e-16
```

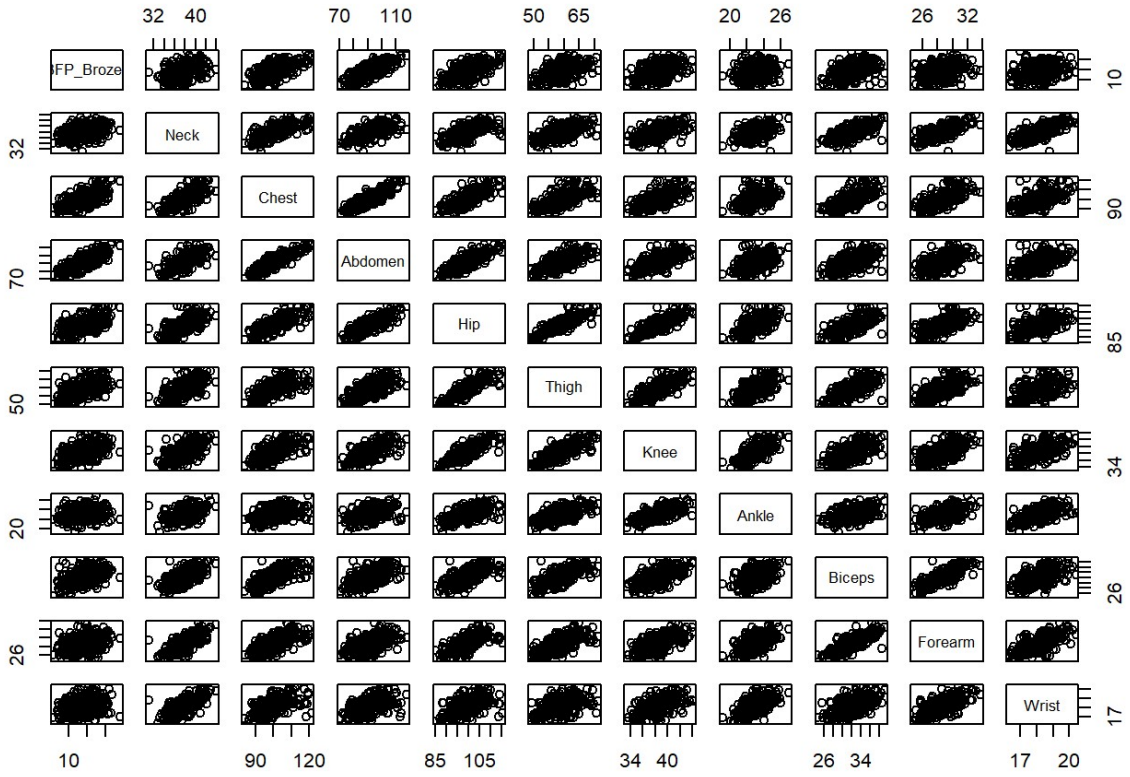
The regression model is $Y = -38.938 + 0.626X$

By observing the p-values for F-statistic and compare them against the 99% significant level (0.01), it could be concluded that there is statistically significant evidence to reject H_0 . In other words, the relationship between abdomen measurement and body fat percentage could effectively be modeled with linear regression.

Since, the objective of this study is to find the best single predictor for estimating body fat percentage, it is essential to model each body circumference measurement against the body fat percentage.

Linear Regression models for each body circumfenrece measurement against the body fat percentage

Visually inspecting the linearity between variables though the scatter plot is a highly effective way of identifying possible linear relationships among the variables. Below scatter plot matrix demonstrates the relationship between each variable in the data-set.



As can be seen in the first column (i.e. scatter plot against the body fat percentage) the Abdomen measurement seems to possess the best linearity due to its narrowness of the cloud.

Additionally, using lm() function in R for each body circumference measurement (similar to the example shown for Abdomen measurement in above), linear regression model parameters have been obtained and shown in below summary matrix.

By considering the R^2 (goodness of fit for linear regression) value for each, the best single predictor for estimating the body fat percentage could be selected. The largest R^2 is evident for Abdomen with 0.642 is selected to be the best predictor. Alternatively, the F-statistic values could also be used in determining the best linear predictor. The high F-statistic value also suggests that the data does not support the null hypothesis.

```
         X Intercept_a Slope_b F.stat R_sqr variability    t_a   t_b
1     Neck     -35.515   1.429  51.09 0.180         18%  -4.69  7.15
2    Chest     -44.997   0.634 174.70 0.429         43%  -9.33 13.22
3  Abdomen     -38.938   0.626 417.70 0.642         64% -13.77 20.44
4      Hip     -54.659   0.737 126.30 0.351         35%  -8.37 11.24
5    Thigh     -30.628   0.832  85.05 0.267         27%  -5.73  9.22
6     Knee     -42.071   1.576  65.20 0.219         22%  -5.59  8.08
7    Ankle      -9.268   1.213  10.99 0.045          5%  -1.10  3.32
8   Biceps     -18.235   1.145  56.18 0.194         19%  -3.70  7.50
9  Forearm     -24.076   1.487  35.34 0.132         13%  -3.35  5.95
10   Wrist     -19.818   2.111  14.58 0.059          6%  -1.97  3.82
           P_a        P_b              CI_a          CI_b
1  0.00000473 0.00000473 [-50.446 -20.584] [1.035 1.822]
2      <2E-16     <2E-16 [-54.499 -35.496] [0.539 0.728]
3      <2E-16     <2E-16  [-44.51 -33.367] [0.566 0.687]
4    5.27E-15     <2E-16 [-67.521 -41.797] [0.607 0.866]
5    3.17E-08     <2E-16 [-41.169 -20.087]  [0.654 1.01]
6    6.22E-08   3.62E-14  [-56.89 -27.252]  [1.191 1.96]
7     0.27155    0.00106     [-25.835 7.3] [0.492 1.934]
8    0.000269   1.37E-12  [-27.945 -8.525] [0.844 1.446]
9    0.000942 0.00000001  [-38.235 -9.917]  [0.994 1.98]
10   0.050099   0.000172   [-39.644 0.009]   [1.022 3.2]
```

Hypothesis testing of α and β for Abdomen to Body fat relationship.

To test the statistical significance of the intercept and slop for the linear relationship between abdomen measurement and body fat percentage, below null and alternative hypothesis are stated:

$H_0 : \alpha = 0$, $H_A : \alpha \neq 0$
$H_0 : \beta = 0$, $H_A : \beta \neq 0$

The 95% CI ranges are obtained using below R code:

```
confint(bfpAbdomenmodel)#95% CI for the fitted line
```

```
                 2.5 %     97.5 %
(Intercept) -44.5099676 -33.366987
Abdomen       0.5660298   0.686797
```

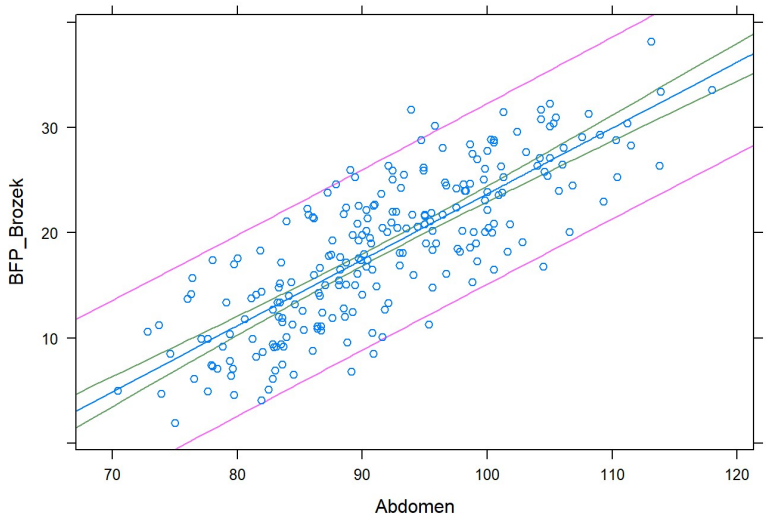
As the 95% CI for α is [-44.51 -33.367] and it is clear that $H_0 : \alpha = 0$ is not getting captured by 95% CI range, hence the results are statistically significant to reject H_0 .

Similarly, as the 95% CI for β is [0.566 0.687] and it is clear that $H_0 : \beta = 0$ is not getting captured by 95% CI range, hence the results are statistically significant to reject H_0 .

Visualising the relationship

Below R code plots the real data points and the linear regression model with some confident interval regions:

```
xyplot(BFP_Brozek~Abdomen,data = body,ylab="BFP_Brozek",xlab="Abdomen",panel=panel.
lmbands)
```



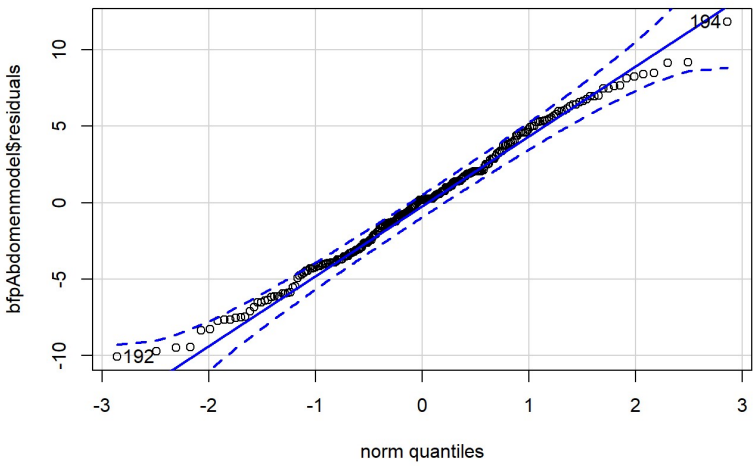
The blue line is the line of best fit for the linear regression. The green bands represent the 95% CI of the body fat percentage mean readings for the regression line. The pink outer lines are the prediction intervals, where the 95% of the data will fall assuming the residuals are normally distributed.

Validating the assumptions

The assumptions of independence and linearity have been validated for the data-set in the previous sections.

The residual value for each data point represent the deviation of the real data from the linear regression model. The existence of the normality of the residuals is important factor in healthy linear regression model. Below R code assess the normality of the residual by plotting the Q-Q plot for residuals:

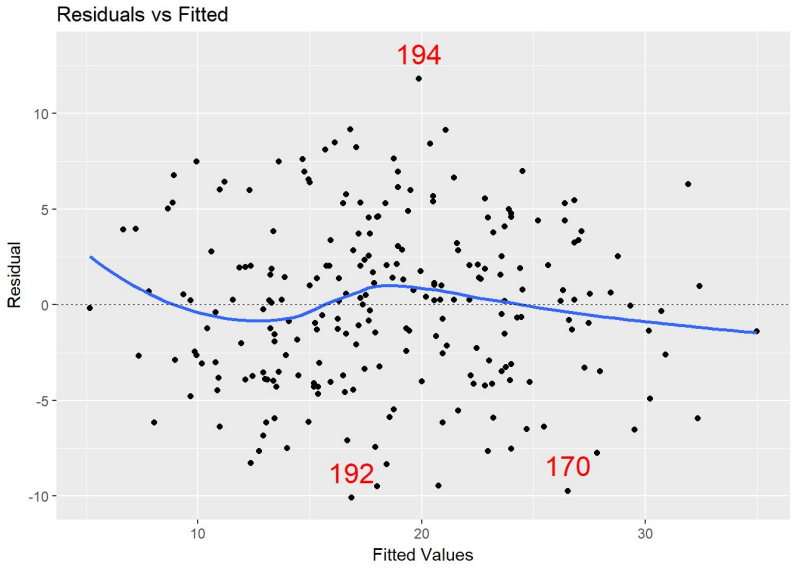
```
qqPlot(bfpAbdomenmodel$residuals,dist="norm") #plot to check the normality of the r
esiduals
```



The plot above suggests there are no major deviations from normality. It would be safe to assume the residuals are at least approximately normally distributed.

The Homoscedasticity is related to the assumption of homogeneity of variance for the two-sample t-test. By plotting the scatter plot for Predicted/Fitted values on the x-axis and the residual values in y-axis, the homoscedasticity for a given regression model can be assessed. Below is the R code for such scatter plot:

```
mplot(bfpAbdomenmodel, 1) #plot to check the homoscedasticity, or constant variance
```



As can be seen in the plot the, the residual value remains the same despite the predictor move across the x-axis. This says the Homoscedasticity assumption used in for this regression model is true and valid.

Conclusion

This analysis has used the body fat percentage data-set of 252 male and female and ten of their key body circumference measurements. The first phase of the study proved that the mean body fat percentages of male and female are statistically close enough to treat them as same. In the second phase of the study, the best predictor for the body fat percentage was identified by comparing the linear regression models of body circumference measurements and body fat percentage in the data-set. The bivariate relationships between body fat percentage and body circumference measurement were inspected using the scatter plot matrix. The 'Abdomen' measurement was selected as the best predictor by comparing the R^2 and "F-stat" of each regression models. Using the best predictor, the human body fat percentage could be estimated by the linear regression equation of,

$$BodyFatPercentage = -38.938 + 0.626 \times Abdomen$$

The positive slope for Abdomen was statistically significant, β = 0.626, and the F-stat values strongly supported the obtained linear regression model. Finally, the visual inspection of the real data points against the curve of linear regression model was utilized to estimate the validity of the regression model. Also the analysis and inspection of residuals supports the normality and homoscedasticity of the data, which further confirms the validity of the assumptions that have been taken in the linear regression modeling.